DOCUMENT RESUME

ED 363 618 TM 020 458

AUTHOR Madaus, George F.
TITLE A Technological and Historical Consideration of Equity Issues Associated with Proposals To Change Our Nation's Testing Policy.
PUB DATE 12 Mar 92
NOTE 50p.; Paper presented at the Ford Symposium on Equity and Educational Testing and Assessment (Washington, DC, March 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Disadvantaged Youth; *Educational Assessment; Educational Change; Educational History; Educational Policy; Educational Technology; Elementary Secondary Education; *Equal Education; Minority Groups; *National Competency Tests; National Programs; *Policy Formation; Student Evaluation; *Test Bias; Testing Problems; Testing Programs
IDENTIFIERS *Authentic Assessment; *High Stakes Tests

ABSTRACT

The equity of a national examination system using authentic assessments is discussed from technological and historical perspectives. Testing is, in fact, a technology embedded in sociotechnical systems such as education, government, and business. The historical perspective is important because it provides lessons for present policy. The first section of the paper offers reasons from the history of technology for proceeding with national testing with caution from an equity point of view. In the second section, the historical evolution of technology is described, and how that evolution brought about the present revolution in testing is traced. Section 3 describes the historical paradox associated with testing policy aimed at assisting the underprivileged. The final section lists the following characteristics of a high-stakes examination system that will be necessary for ensuring equity: (1) a level playing field of social and educational conditions; (2) a clear definition of purpose; (3) the need to recruit more minorities into testing; and (4) the need for an independent monitoring agency. The technical and equity considerations of a national examination system require that we proceed with the utmost caution. (Contains 115 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY
GEORGE F. MADAUS
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)"

ED 363 618

A Technological and Historical Consideration of Equity Issues Associated With Proposals to Change Our Nation's Testing Policy[1]

by

George F. Madaus

Boston College
Chestnut Hill MA 02167

Paper Prepared For The Symposium On
Equity and Educational Testing and Assessment

Sponsored by the Ford Foundation

Washington D.C. March 12, 1992

TM020458

BEST COPY AVAILABLE

American testing is in tumult. Traditional, standardized, norm-referenced, multiple choice tests -- the policy choice for the past 70 years -- are under attack. The most widely touted policy alternative is a national system of examinations which would employ "authentic" assessment techniques. These techniques, do not require students to select their answers from given alternatives; rather they require students to construct answers, perform, or produce something for evaluation. Authentic assessments, it is asserted, will be worth teaching to; will defeat negative test preparation effects associated with multiple choice tests; will give teachers clear models of acceptable outcomes; will have a positive influence on instruction and learning; will measure higher order skills; and will lay bare examinees' thinking processes. The unmotivated will be motivated, all students will meet "world class" standards, our nation's productivity will be increased, and our global competitiveness restored. But more importantly from our point of view here today, proponents claim that the examination system will help to improve the academic performance of all students, regardless of race, home language, culture, or gender; further, the assessments will not treat individuals or groups unfairly. As the new administration begins to formulate its policy agenda for elementary and secondary education it is important to submit these claims -- and particularly those related to equity -- to a relentless and thorough examination of every reasonable technical, educational, societal, and policy difficulty or objection.

I have been asked to discuss the equity of a national examination system using authentic assessments from two perspectives --

technological and historical. Why these two perspectives? First, testing is a technology, and so should be considered from the general perspective of the effects of technology on institutions and individuals, and technology's relationship to equity. Testing is so ubiquitous, taken for granted, entrenched as a tradition, that people do not generally think of it as a *social technology* and *a technical craft*. Nonetheless, this is precisely what testing is: a technology embedded in such sociotechnical systems as education, government, and business.[2] Given that testing is a technology, filtering a new testing program through our knowledge of the impact of technology in general helps to illuminate testing issues in a new way.

I am not a latter day Luddite opposing technology or new testing proposals. Technology, of course, has powerful positive aspects.[3] It gives people the power and will to create new possibilities and in turn, more choices, more opportunities, and more freedom. It can enhance human dignity, and aid human aspirations (Applebaum, 1992). Thus, the benefits of technology are enormous, and, we simply cannot do without sensible testing programs. Nonetheless, my point today is to remind you of something you already know. As technology affords new opportunities, it simultaneously creates problems, opens new ways to make big mistakes, alters institutions in unanticipated negative ways, and impacts negatively on certain populations.[4]

The historical perspective is important, not because history repeats itself, but because, as Elton puts, it history teaches you to "think more deeply, more completely and on the basis of an enormously enlarged experience about what it may be possible or desirable to do now." In terms of the present debate he points out

that "a knowledge of the past should arm [people] against surrendering to the panaceas peddled by too many myth makers" (Elton, 1991, p.72-73).

While history may not repeat itself, I believe that human nature and the motives and actions that flow from it are ageless. Therefore, the ways in which past testing programs have been corrupted, or have come to disadvantage various populations through the actions of key players should be a cautionary tale in the contemporary policy debate. To think that currently proposed assessment techniques are immune from producing a disproportionate negative impact on minorities, or from the misuse and corruption associated with essay or multiple choice high stakes tests over the last hundred years or more is to be simply oblivious to the strong pull of self-interest.[5]

In the first section of the paper I offer reasons from the history of technology why from an equity point of view we should proceed with caution. In the second section I describe the historical evolution of the technology of testing and how that evolution brought about the present revolution in testing. Section 3 describes the historical paradox associated with testing policy aimed at assisting the underprivileged. In the final section I enumerate the characteristics of a high-stakes examination system that are necessary for ensuing equity.

Before beginning I need to clarify some terms. First, I focus on assessment exclusively in its role as a large-scale, high-stakes, policy tool to drive reform and make important decisions about individuals.[6] Second, the word assessment, modified by the adjectives new, authentic or alternative, has become an approbation, while the term

standardized, norm-referenced, multiple choice test is greeted with intolerance and disdain. The term assessment has captured the linguistic high ground much as the term minimum competency testing did in the last decade. The positive images conveyed by these terms mask their many functions and side effects. Hottois (1987) has argued that technology -- and I would add testing in particular -- "flourishes especially well under the mask of humnism" (p 79). But strip away the linguistic veneer, and whatever noun you chose -- assessment, exhibitions, examinations, portfolios or just plain test -- they all rest on the same basic technology. That is, you elicit a small sample of behavior from a larger domain of interest, e.g., Algebra or Aptitude, to make inferences about the person's probable performance relative to the domain. And, on the basis of the inference, you describe or make decisions about individuals or institutions. The equity issues embedded in this deceptively simple technology relate to preconditions that influence the person's or group's ability to exhibit the behaviors sought in the first place, and then to the nature of the specific techniques used to elicit them. In the final analysis, assessment equity issues come down to the question of how correct the inferences and the concomitant descriptions or decisions are for different populations.

1. THE TECHNOLOGICAL PERSPECTIVE

The danger of enchantment with technological solutions

Because of the strong belief in the Western world in the "religion of progress," technological solutions are likely to be accepted as good simply because they represent progress (Winner

1986, p. 170). People are captivated by the latest thing, the latest gadgets and latest instruments of power, by results and immediate consequences of using devices (Ellul, 1992; Merton, 1964). Popular ideology has reified technology, so that technological images have overshadowed function and effects (Frost, 1992). Over the past forty years educational reform is a case in point, and the fascination with authentic assessment and a system of national exams is a continued manifestation of this faith. But we need to step back and ask if there is any evidence that the introduction of assessment procedures will solve the problems it is proposed to solve, or whether we should be looking for solutions elsewhere (See Sarason, 1991).

The danger is that the prestige and aura surrounding technological solutions will lead us to pursue a particular technical solution when something quite different may be called for (Taylor, 1991).[7] Taylor points out that the approach in medicine not to treat the whole person with a life story, but rather to see the individual as the locus of a technical problem, has led to the undervaluing of nurses who more often give humanly sensitive care as against that of the specialists with high technical expertise (Taylor, 1991). Technology has also altered the relationship between physician and patient by increasing alienation between them as well as altering the system of health care delivery and its costs (Reiser, 1978).

In the case of alternative assessment, we may be so captivated by its status as a technology that we may too readily accept it as the solution to the perceived problems of underattainment in our schools. In fact, something quite different may be called for such as, attending to student health, nutrition and living conditions, changing teacher

pre- and in-service training, instituting a system of school inspection. There is a real danger that our love of the technical solution will blind us to the reality that we cannot test our way out of our educational problems (National Commission on Testing and Public Policy, 1990). As Sizer puts it, reforms like testing do not go "to the heart of the problem: the fundamental misdesign of schools and the instability of the families and communities from which students come" (1984, p. xiii) Further, like Taylor's medical example, assessment as a high-stakes policy tool undervalues the daily, humanly sensitive interaction between teachers and their students in the complex social system of the classroom-- the key to solving our educational problems.

Unintended effects associated with technological solutions

In considering the implementation of a new testing policy we need to bear in mind that technologies may have profound effects beyond their designed intent; they typically perform an indefinitely large number of further unanticipated functions, produce a large number of effects, and project a large number of meanings and symbolic values (Eclove, 1992). The history of technology is replete with examples of specific technologies changing society in totally unexpected ways; consider the clock, automobile, central heating or indoor plumbing, or the rail road.[8] Rather than serving consciously chosen, limited social ends, technology autonomously shapes and reshapes social life according to its needs (Ellul, 1964, 1992; Fielder, 1992; Winner, 1977). Institutions, individuals, attitudes, and ideas change to accommodate technology (Fielder, 1992). One commentator laments, "Like the sorcerer's apprentice, we have

brought technology into being to serve us and instead find ourselves its servants" (Fielder, 1992 p. 107).

This phenomenon of unintended effects is borne out by the history of the technology of testing. For example, the requirements of testing and assessment in the policy sphere over the past 150 years reshaped selection procedures, curriculum, attitudes about the ends of education, and created bureaucracies and other ancillary commercial institutions such as testing companies and cramming schools.[9] Results from a test designed for one purpose were employed by many different users for many alternative purposes. Test results differentially affected different individuals and groups through tracking, classification, selection and placement, and the denial of entry-level opportunities in the workplace. Test results have become equated with the ends of education (National Commission on Testing and Public Policy, 1990).[10] Finally, much of the history of test use in the policy arena is an ill-disguised attempt to coerce the actions of teachers or students or both, to alter the instructional delivery system by linking results to high-stakes. And there is no doubt that such pressure works. But it does so at the price of corrupting first instruction and then the very tests themselves.

Technology and equity

Technology is not by its nature socially unjust. It may in fact have a weak tendency toward equity (Borgmann, 1984).[11] However, technologies are shaped and influenced by a range of social, professional, political, economic, technical and cultural factors that already exist in society. As a product of the existing structure of

opportunities and constraints in society, a particular technology is likely to extend, shape, rework, or reproduce that structure in ways that may be more or less unpredictable but that mirror the forces that shaped it (Bijker & Law, 1992).

Inequity associated with a technology may be difficult to detect since most technologies are based on highly technical, arcane underpinnings. Indeed, some modern technologies are so opaque that we do not think of them as technologies at all. And because of this they do there work for good or ill without much criticism or awareness (Postman 1992). For example, consider credit checks, bank loans, the triggering of an IRS audit, econometric modelling, or census projections; all of these are dependent on the invisible and abstract technology of computer modelling and computerized decision procedures (Wallace, in press).[12] These commercial, financial, modelling services at their heart, are, like test scoring and scaling, based on hidden algorithms. Wallace, in words applicable to testing, reminds us that there are at present few safeguards for the public against deliberate and unintended misuse of data generated from such hidden abstract algorithms.[13]

More importantly, from the point of view of equity, many people simply are not aware of the control these hidden systems can exert on their lives. Nor are they cognizant of the biases, values, and assumptions held or made by the technical elite who crafted them, and who might have failed to take into account all the groups who would be affected by a model's results (Wallace, in press; Winner, 1992b).[14]

Americans usually do not stop to inquire whether a test or any other technology might have been designed and built in such a way that it produces a set of consequences or inequities logically and temporally prior to any of its professed uses (Winner, 1986). For any given technology its functions follow from its form (Postman1992). Langdon Winner describes the unintentional way in which inequities arise with the adoption of a particular technology. He writes:

> [There] are instances in which the very process of technical development is so thoroughly biased in a particular direction that it regularly produces results heralded as wonderful breakthroughs by some social interests and crushing setbacks by others. In such cases it is neither correct nor insightful to say "Someone intended to do somebody else harm." Rather, one must say that the technological deck has been stacked in advance to favor certain social interests and that some people were bound to receive a better hand than others. (Winner, 1986 p. 25, emphasis in original)

The entire test development process[15] -- the constructs we measure, the cultural background and specialized training of test developers, the material chosen for inclusion and exclusion, the design choice of individual items (e.g., the multiple choice, extended essay format, performance, etc.)[16], the language and idioms used in the questions, the directions given to administrators and examinees, the validation process -- all serve to stack the testing deck in favor of certain values and groups in our society, and unintentionally assure that other values and groups are dealt a weaker hand prior to any actual use of the test.

Likewise, the decision to use a test for one purpose rather than other also unintentionally stacks the deck against certain individuals and groups.

Values that underlie much of testing used in the policy sphere include utilitarianism, economic competitiveness, technological optimism, objectivity, bureaucratic control and accountability, numerical precision, efficiency, standardization, and conformity. In turn, testing shapes important educational values -- for example, conceptions of quality; student and teacher-reward systems; what an educated person should know, understand, and be able to do; what should be taught and learned as well as how it should be taught and learned.

Since social and cultural groups differ in the extent to which they share the values which underlie testing and the values which testing promotes, it is clear that the imposition of a testing system raises questions relating to equity. For the most part, the predominant values associated with testing, though they may not be explicit, have been those of policymakers, test users and the technical community. To deal with equity problems that will inevitably arise in this situation, it is important that the values of testing be explicated and examined in the context of different populations of test takers and other consumers.[17] The views of different races and cultures, and of those with different home languages, as well as gender considerations, cannot be carefully marginalized, pushed to the periphery, and remain largely unexamined; instead they must be weighed against the predominant values embodied in present testing policy (Ellul, 1992).

As Postman notes, "the benefits and deficits of a new technology are not distributed equally. There are, as it were, winners and losers" (1992, p. 9). We need to constantly ask who wins and who loses if we introduce a new national high-stakes assessment system; whose and what values will be promoted, which groups will benefit, and which will be dealt the weaker hand by putting such a system in place? In particular, given that institutions and students will be treated differently according to their performance, the following questions need to be addressed:

- Will current proposals that call for assessments at at least three grade levels linked to "world class standards" for all students in five subjects shrink the range of school subjects that society values to those most closely related to academic aspirations at the expense of other career aspirations, skills, and interests?
- How will non-academic non-college bound, or at-risk students perceive the imposition of a common core of five subjects with common world class standards?
- What specifically will happen to a fourth, eighth, or twelfth grader who scores poorly on a national examination -- how will he or she be treated?
- Will a national system of examinations eventually become linked to vouchers and a market driven system of education? And if so what are the equity, and educational implications of such use?[18]

Technology, power and politicization

All technologies should be viewed, "as a strategy involved with the larger structures of power in society, as a part of the culture and

social process of particular societies" (Applebaum, 1992 p. 528).[19] Policymakers and the technical elite are aware of this important fact.[20] Technology (and testing) is politicized for two reasons:[21] first, because of the intrinsic capacities or powers it makes available when it is used, and second, because of the power that the deployment of the technology affords those who control it (Wartofsky, 1992).[22] We must distinguish between the technological power that new assessments provide, in terms of *potential* information (direct performance, higher order skills etc.[23]) -- that is technical power -- and the power which control over such an information source would give those in control of the system at the expense of local districts, schools and teachers -- that is the political power of testing programs.

In testing, effective policy decisions are made by governmental bodies at federal and state levels hierarchically controlled, in conversation with technical experts.[24] There are two dangers here. First, technical truth can easily become hostage to political power in the decision making procedure. I have witnessed a tendency to dismiss reasonable conjectures about technical problems of new assessments with the proposal that we build the plane while we fly it. Advocates promise that technical and equity issues will be addressed when they arise, and ask us to trust them to solve them.

Second, the public may have, at best, only indirect input into such decisions. The widespread technological optimism noted above leading to the adoption of technological solutions to society's problems, has contributed to an erosion of citizen participation in decision making around issues of technology.[25] However, if we believe that all citizens have a right to be directly involved in

technological choices that affect them -- such as the creation of a national system of exams -- then means must be found to realize that right in practice (Winner, 1992a).[26]

Contemporary assessments of technological risks typically employ decision rules that take little or no account of the possible damage to the lay persons who are most likely to be harmed by a risky technology; also overlooked is the value-ladenness of certain of risk-cost-benefit analysis interpretations (Shrader-Frechette, 1987, 1992). Arguably, the risk, particularly for certain groups of test takers (e.g., minorities, women, handicapped) in certain contexts, (e.g., high-stakes uses of results), are high. And even the best constructed test can be subject to misuse and abuse by a diverse array of users. Nonetheless, risk, abuse and misuse are not among the first words to come to mind when most people, especially policymakers, think about testing. The perils and potential for abuse inherent in testing are simply not as dramatic or as obvious as those associated with other products or activities. There is an insufficient level of *generalized* outrage to prompt policy makers to take effective action in overseeing the testing enterprise. I will return to this point in the final section (Madaus, Haney, Newton, & Kreitzer, 1992). There is an irony in testing. People are defined by it, and aspects of society are altered in unintended ways-- often in negative or limiting ways -- nonetheless, there is a relentless endorsement of its use without proper safeguards to make very important decisions.

2. THE EVOLUTION OF THE TECHNOLOGY OF TESTING

The evolution of testing technology as a major administrative mechanism in the policy arena sheds light on the problems associated with implementing a high-stakes authentic assessment policy today (Madaus, in press). Like all technologies, testing evolved through stages of increasing complexity occasioned by expanding use.[27]

In the case of any technology some components fall behind others, acting as a drag on the entire system by functioning less efficiently, malfunctioning, or adding disproportionately to costs (Hughes, 1989). Testing provides an excellent example of this phenomenon. The history of changes in testing in Europe and America is the story of eliminating features of testing that acted as a drag on the system, particularly as the number of examinees increased.

Changes over the last two centuries in the predominant way of examining student achievement -- from the oral mode, to the written essay, to the short answer form, to the multiple choice and then to machine scorable answer sheets, and finally to computer adaptive testing -- were all geared to increasing efficiency, and making the assessment system more manageable, standardized, easily administered, objective, reliable, comparable, and inexpensive as the numbers of examinees increased.[28] These technical evolutions eventually made feasible the federal and state legislation of the last forty years mandating the large scale district and state multiple choice testing programs.

The end result in this technical evolution of testing -- a highly economically efficient, standardized test -- is analogous to the

invention of the mechanical tomato harvester in the late 1940s. These machines are "able to harvest tomatoes in a single pass through a row, cutting the plants from the ground, shaking the fruit loose, and (in the newest models) sorting the tomatoes electronically into large plastic gondolas that hold up to twenty-five tons of produce headed for canning factories" (Winner, 1986 p. 26). However, such efficiency came at the cost of introducing new varieties of tomatoes that are hardier, sturdier, and certainly less tasty than those previously grown. It also had an unanticipated but profound impact on rural agricultural communities (Winner, 1986).

Our current "hardier, sturdier but less tasty" tests, with their negative effects on instruction, are presently seen by advocates of "authentic" assessment as a serious drag on the reform process. I believe that the evidence supports this view; there is little doubt that the high-stakes multiple-choice testing programs of the 70's and 80's did not measure the many important cognitive skills now called for by reformers, and have had an overall deleterious effect on teaching and learning (Darling-Hammond, 1991; Jaeger, 1991; McLauglin, 1991; Shepard, 1991a; Stake, 1991). Recent research by my colleagues and I on standardized tests in math and science showed that these tests, while technically state-of-the-art, were overwhelmingly measuring lower order thinking skills. They were not tapping the domain of skill and knowledge presently demanded by the math and science communities. Moreover, our national survey of teachers revealed that those with greater than sixty percent minority pupils in their classroom were far more apt to teach to the standardized tests -- with their emphasis on lower order skills -- engage in various kinds of

direct test preparation activities, and spend more precious class time in such pursuits, than were their peers in predominately white classrooms (Madaus, West, Harmon, Lomax, & Viator, 1992).[29]

A major premise of the movement towards a national examination system or systems is that if the form of assessment is changed, then its effects on learning will differ from the effects of the old multiple-choice systems, even though other conditions associated with the tests, particularly their external control, the attachment of high stakes, and their use to drive reform, remain the same. Research relating to the effects of the form of measurement on individuals and groups is not very extensive. It does indicate that the method of measurement contributes to variance in measured achievement, but it does not consistently verify particular methods as conferring advantages for any particular groups.

As far as the impact of performance assessment on linguistic, racial and cultural minority groups, a review encompassing all the major data bases turned up very little information.[30] The available information does nothing to encourage the belief that alternative or performance assessments will be more equitable than other forms of assessment. For example, an examination of performance on NAEP open-ended essay exams and multiple choice reading tests show that achievement differences between African-American and white students are the same regardless of test type (Baker, O'Neil, & Linn, 1991; Linn, Baker, & Dunbar, 1991).[31] Further, there is evidence that African-American and female examinees perform better on computational math items than on contextualized word problems (Hambleton & Murphy, 1992). Again, LeMahieu found that African

Americans received lower scores on their portfolio evaluations than whites regardless of race of rater. On another long term independent record of writing, more than 70% of those classified as high writers were white while more than 80% of low writers were black. When the writing from the portfolios was compared with this independent record, the high writers on the independent measure scored even higher on the portfolio samples. The difference seemed to be the self selection of materials. The African-Americans did not choose material from their portfolios that presented their best writing (LeMahieu, 1992).

The paucity of hard data on this issue inevitably leads to the conclusion that claims of proponents that authentic assessment will narrow the achievement gap simply lack supporting evidence. This is a strong reason to proceed slowly along the research rather than implementation path.

Performance-based assessments may well be preferable to traditional tests in terms of their effects on the ways many teachers and students spend their time and the aspects of the curriculum to which they pay attention. Such assessments may even eventually result in improved performance for minority children. However, when we come to anticipate the effects of their use in high-stakes situations, we must consider the experiences of the high-stakes testing programs of the 80s that corrupted instruction and the tests themselves (Darling-Hammond, 1991; Jaeger, 1991; Madaus & Kellaghan, 1992; Madaus, 1991; McLauglin, 1991; Shepard, 1991b; Stake, 1991). We have no evidence to believe that performance-based measures will not be as corruptible as any multiple choice measure

when used in the context of measurement driven instruction (Madaus, in press).

There are further problems relating to the use of performance-based assessment that have long been realized in the history of testing. Among such problems are: inefficiency; administrative inconvenience; subjectivity and bias; the amount of time required when large numbers of pupils have to be assessed; difficulties in standardizing conditions of support for teachers within a school who are administering them, and for the actual administration itself; lack of comparability of results; poor generalizability; and high cost (Dunbar, Koretz, & Hoover, 1991; Koretz, Linn, Dunbar, & Shepard, 1991; Madaus, 1991; Madaus & Kellaghan, 1991a&b, 1993; Nuttall, 1992; Shavelson, Baxter, & Pine, 1991, 1992).

These problems and their attendant equity ramifications lead me to conclude that setting up a national system of examinations as a large-scale policy tool for certification of individual students is premature. We need much more research and development before embarking on yet another high-stakes testing adventure.[32]

3. HISTORY OF THE PARADOX OF TESTING PROGRAMS AND UNINTENDED EFFECTS PARTICULARLY ON MINORITIES.

Technology leads a double life. One conforms to the intentions of policymakers and designers; the second contradicts them, proceeding behind their backs to produce unanticipated uses and consequences (Noble, 1984).[33] As Postman (1992) shwerdly observes, "[M]ost people believe that technology is a staunch friend.. It is the kind of friend that asks for trust and obedience, which most people

are inclined to give because its gifts are truly bountiful. But, of course, there is a dark side to this friend. Its gifts are not without a heavy cost.. Technology, in sum, is both friend and enemy" (p.xii).

The history of testing over the past 200 years is a classic example of this double life, of this friend/enemy relationship. On the one hand, the intent of policymakers in mandating testing programs has been corrective rather than malevolent, or punitive. Policymakers have seen examinations, coupled with rewards or sanctions, as a practical, efficient and cheap means of correcting what they consider to be defects in the educational system.[34] Thus, tests have been mandated for such laudatory purposes as eliminating patronage; opening access to groups hitherto excluded; assuring basic literacy and numeracy for all students; establishing and/or maintaining standards of excellence; holding teachers, students or schools accountable for learning; and allocating scarce resources to the most able or deserving.

On the other hand in using tests to address practical problems, abstract principles about test validity and unintended negative consequences seemed remote. Unfortunately, the well intentioned testing programs often took an ironical turn. Testing programs, designed as corrective, often resulted in unintended negative consequences, which though unforeseen by their creators, impacted heavily on the schools and on certain populations.[35]

Let me first offer a general example of this paradox and then briefly three specific examples.

• Testing and the rise of individualism.

The use of tests in policy contributed in *small* measure at least to the rise of individualism, which, it can be argued, is one of our

greatest achievements, giving people the freedom to be themselves (Taylor, 1991).[36] Test results were used extensively by policymakers in the 19th century for the purpose of replacing status, connections or place in the community with individual merit as the criterion for educational and occupational selection.[37]

However, this policy simultaneously contributed to the rise of the bureaucratization of education,[38] to the factory model of thinking about schools,[39] and, most importantly, to the classification of people by a numerical score.[40] The quantified mark for answers to individual questions, introduced in the latter part of the 18th century, made it possible to control the actions of the examinees as well as school personnel (Hoskins, 1968).

Tests became one tool in the rise of the credo of instrumental reasoning -- the rationality of efficiency.[41] Students had to periodically reveal how their learning was progressing, and the exam defined what was expected. The examinee for the first time became a "describable, analyzable object" and the examination the "ceremony of this objectification" (Foucault, 1979 p 190). Further, it was now possible to accumulate student marks, organize them, rank them, classify them, form categories and groups, determine averages and fix norms.[42] Thus, we find the paradox of using merit as a criterion of individual selection, while simultaneously enhancing bureaucratic control over individuals and groups through the reductionism of quantification. In terms of equity, this quantitative reductionism has over the past century impacted most heavily in a negative way on minorities.

Let me briefly offer three more specific examples from many possible historical and contemporary programs of the paradoxical aspect of testing policy.

• Binet's development of the first individually administered IQ

Binet's work on the measurement of intelligence originally had a positive practical purpose. Parisian policy makers needed an objective means to identify children unlikely to succeed in normal classes and therefore in need of special instruction. A laudatory goal! Nonetheless, there is little need to relate the story of how IQ testing evolved differently, and became a defining technology, one that influenced institutions and attitudes in profoundly negative ways (Gould, 1981, Postman, 1992).[43] The basic technology Binet developed, subsequently made more efficient by Otis, was there, and because it was, it was used to misclassify and label people through most of this century. The IQ technology, like all technologies, is indifferent to the use made of it. Use is extraneous to the technique itself but tendencies relative to use are often embodied in the technique's social existence (Simon, 1983). People yielded to the temptation of interpreting a test score as a "true" presentation of reality. The test has transformed the multifaceted, abstract concept intelligence into "a technical and exact term that leaves out everything of importance" (Postman, 1992,p. 89) This in turn help concretize a racist, eugenic, view of intelligence.[44] Clearly, the unintended negative outcomes brought about by the widespread policy use of IQ tests disproportionately disadvantaged minority populations. For example, it is ironic that despite Binet's original purpose to identify children in need of instructional assistance, the IQ test in this country

led to blacks and Hispanics being disproportionately placed in dead-end classes for the educable mentally retarded.[45]

• Chapter 1 testing

The original testing provisions of the now Chapter 1 (then Title 1) were originally put in place in the the 60s for legitimate accountability purposes to obtain evidence that programs for disadvantaged students were working. However, the regulations quickly lead to the domination of low level multiple choice tests being used for important programmatic decisions. Further, the power of those tests contributed to the corruption of instruction for poor children and eventually to the corruption of the test results themselves (Cannell, 1987, 1989; Linn, et al., 1989, Commission on Chapter 1, 1992).[46]

• Minimum competency testing

As a final example, consider the minimum competency testing movement. Policy makers originally mandated such tests to ensure that all students had achieved basic numeracy and literacy skills. Ironically, however, it led to the corruption of instruction through a heavy emphasis on low level skills; a subsequent rise in test scores without concomitant improvement in skills that the tests were originally designed to measure, and the introduction of test preparation techniques and materials to train students to take them (Haney, et al., 1993 forthcoming; Koretz, et al., 1991; Madaus & Kellaghan, 1992; Madaus & McDonagh, 1979). Further, minorities disproportionately failed these tests and were subsequently subject to more intensive test preparation activities geared at low level skills (Haney, et al., 1993 forthcoming; Koretz, et al., 1991; Madaus &

Kellaghan, 1992; Madaus & McDonagh, 1979; National Commission on Testing and Public Policy, 1990; Madaus, et al., 1992).

In considering a major change in our national testing policy we need to keep the paradox associated with past policy in mind. We need to look at history and ask, "What are the likely unintended outcomes that will likely result from the proposed national system of exams using alternative assessments, how will they affect on groups traditionally disadvantaged by tests, and how can we minimize these negative effects on minorities in particular and on education in general?"

4. WHAT MIGHT AN EQUITABLE NATIONAL TESTING PROGRAM LOOK LIKE?

Despite the potential problems and unintended outcomes that are likely to be associated with the proposed new testing policy, no one would deny the need for information on how our schools are doing. Policymakers have an obligation to collect accountability information and the public has a right to know. In this last section I sketch what I see to be the essential preconditions that would help policymakers craft a more equitable system of national examinations; one that might minimize the negative kind of side effects associated with high-stakes policy-driven testing programs in the past.

• Need for a level playing field

Equity issues arising from assessment are not confined to the specific instrumentation, technique, or psychometrics inherent in the test itself. More broadly they relate to antecedent social and

educational conditions, and to the unintended consequences different uses may reasonably be expected to produce. Elsewhere I have described how Witold Kula, in his fascinating treatment of the history of weights, measures and standards, describes the eternally cherished dream of humanity for the "just" measure (Madaus, in press). Kula recounts that in the Hebrew Bible, the New Testament and the Koran, realistic social conceptions of measures were transformed into the symbolic "just" measure -- a symbol for the just person, for justice as such, and of just human relations (Kula, 1986). It is precisely this symbolic concept of the "just" measure that needs to be in the forefront of any analysis of creating a national system of exams.

As Kula documents, attempts to standardize and make weights and measures equitable and just failed until other concomitant reforms in such things as feudal rights, provincial particularism and equity before the law were set right (Kula, 1986). Likewise, I believe that any proposed exam system simply cannot produce truly "just" measures until policymakers put in place appropriate national delivery standards for social, health, family, and educational resources and support systems. Therefore, in addressing the equity of alternative assessments in a high-stakes policy-driven exam system we need to create first and foremost a level playing field for students and schools.[47] Only then can we begin to claim that a national examination system is an equitable technology for making decisions about individuals, schools or districts.

The tremendous value of the National Council on Education Standards and Testing (NCEST) report is that it raised the politically touchy issue of delivery standards (National Council on Education

Standards and Testing, 1992). But delivery standards aren't enough. We also need national systems that implement those delivery standards equitably across states and districts, and within districts across schools. It is ironic that the creation of an examination system geared to "world class" standards is moving ahead at a much faster rate than is the development of delivery systems to help to level the educational playing field for all students. We cannot assess our way out of our educational problems. A new testing program (whatever its form) will not help children described in Kozol's *Savage Inequalities*, unless and until other systemic reforms are put in place (Kozol, 1991).

• Need for clear definition of purpose of the assessment system

A perennial problem of testing programs is that policymakers and others wish to use a single instrument for a multitude of purposes --- e.g., to foster good teaching and learning, to make high-stakes decisions about individuals, to hold schools and districts accountable, to facilitate a voucher system, and to monitor national progress toward realizing the President's and governors' educational goals. Long experience with issues of test design, scoring, reporting, and the needed supporting infrastructure indicates that these different purposes require different procedures and techniques.[48]

It seems to me there are two possible purposes the new system of exams might satisfy. The first is to provide policymakers information for national, state, district or school level accountability. The second is to provide information to make high-stakes exit, certification, or entrance decisions about students. Whichever of the two options is chosen, the system of assessments must be geared to

well defined and articulated curricula which needs to precede the assessments -- not arise out of them. My bias is that right now the former is the more equitable and most feasible alternative. Further, it supplies federal and state policymakers with the proper level of information they need (Madaus & Kellaghan, 1993; Madaus, in press).[49] Programmatic, institutional or system accountability -- can be satisfied with a NAEP-like assessment system involving sampling of tasks and students. Matrix sampling of assessment tasks across samples of students cannot produce individual scores, but it can, if done correctly, yield rich institutional data.[50] One of the points in Deming's total quality management model is that we should cease dependence on mass inspection; "routine 100% inspection to improve quality is equivalent to planning for defects" (Deming,1982, p 28).

The unavailability of individual scores under this plan means that the assessment cannot be used for tracking, certifying, or classifying individual students. This in turn lowers the stakes, at least from the individual student's point of view, and therefore, is more equitable right now for minorities. Further, NAEP like sampling is a cost effective way of using performance and product assessments which are considerably more expensive that traditional tests (Madaus & Kellaghan, 1991a&b; Madaus & Kellaghan, 1993; Popham, 1993). Finally, this plan overcomes as yet unsolved problems of comparability, of standardization of tasks within and across schools -- including among other things administration, antecedent instruction and ancillary skills-- and of generalizability of sample tasks to larger domains, all problems associated with the use of alternative

assessment techniques to make high-stakes decisions about individuals (Madaus & Kellaghan, 1993; Nuttall, 1992).[51]

If policy makers choose the second alternative -- a system for high-stakes individual decisions -- they should do so only after the other systemic reforms referred to above are in place, and the myriad of technical and equity problems associated with this alternative are solved. They also might look more closely at the European and old New York State Regents examinations. Both these systems use(d) a combination of essay and multiple choice formats, with the essay mode predominating in Europe.[52] These modes of assessing not only make this choice more feasible economically, but they are also less burdened by as many of the technical problems as those associated with alternative assessment techniques.[53] Further, there would be no need to create an entirely new examination infrastructure. With proper time and modifications, the present national structures of the CEEB or ACT could serve both developmental and delivery functions.

If the second option is pursued, then I feel the following should happen:

- It should be phased in very slowly and be subject to careful, independent monitoring.
- It should be instituted only at ages 16+ and 18+ as is the case in our competitor countries (Madaus & Kellaghan, 1991a&b).[54] It should not be implemented at the primary levels where the possibility of tracking and retention are high. We can obtain accountability information about the elementary sector by other means.

- It should encompass many more than the five subjects mentioned in the National Goals. In Europe for equity reasons, the number, and level of exams available to students to choose from was greatly extended to 30 or more subjects so that those who were not university bound could leave secondary school with some credential (Madaus & Kellaghan, 1991a&b).
- It should be as decentralized as possible to help attain "local" technological flexibility and pluralism and to empower teachers and to demonstrate respect for their judgments about students.
- It needs to incorporate extensive in-service training. Alternative assessments in a high-stakes context are not well enough understood, resourced, supported or linked to curriculum and current teaching practice to be implemented conscientiously by teachers without proper preparation. It is much easier to plan and mandate a new assessment program than it is to change attitudes and the institutional practices of people involved in the day to day life of classrooms.

• Need to recruit more minorities into testing

Currently there are far too few persons of color working in the testing field. Their experience and insight are needed when a new alternative assessment system is being designed. We need financial support and other incentives to encourage minorities to enter the field of testing. Training programs should focus not only on technical skills but also concentrate on wide-ranging, non-technical aspects of the testing enterprise, including social, political, ethical, legal, and economic issues (National Commission on Testing and Public Policy,

1990 p. 31) One of the reasons that gender issues in testing are finally receiving the attention they have long deserved is because of the critical mass of women now working in the field.

- **Need to establish an independent agency to monitor the testing used in the policy sphere.**

In 1990 the National Commission on Testing and Public Policy recognized "the need for sound, fair, and reasonably efficient mechanisms to help make difficult decisions about individuals and institutions" and called for "the use of tests and assessments to open the gates of opportunity for America's diverse people"(National Commission on Testing and Public Policy, 1990 pp. 1-2). The Commission concluded that to realize this goal, "The enterprise of testing must be subjected to greater public accountability;" further, they recommended "the development of additional institutional means to examine the quality of tests and assessment instruments and to provide oversight of test use" (National Commission on Testing and Public Policy, 1990 p. 32).

The nation has never had such institutional means to monitor testing programs. Policymakers and test users can turn to an extensive commercial and governmental infrastructure that has evolved over the past 90 years to assist them in test development, administration, scoring, and reporting. But there is no analogous infrastructure to which they can turn to evaluate independently a testing program before adoption, or to monitor test use and impact after implementation.

The need for an institutionalized means to perform such functions seems all the more urgent given the very real movement toward a national test or testing system.[55] The benefits and hazards to institutions and individuals associated with the creation of a national examination system are potentially enormous. What other institution in our society would contemplate such a dramatic treatment without first evaluating claims in the light of what we already know? And then if deciding to proceed, would only do so slowly and systematically, with some independent mechanism to monitor the consequences?[56] The creation of an independent agency to monitor any national system of examinations paid for by taxpayers seems to me to be badly needed.

Elsewhere, my colleagues and I have detailed what such an organization might look like, what it might do, how it might operate, and how it might address equity concerns (Madaus, et al., 1992). Our conclusions were based on a three year study supported by the Carnegie Corporation and the Ford Foundation.[57] The proposed independent organization would go a long way towards addressing equity concerns and ensuring that the consequences associated with the implementation of a national system of exams are continually monitored and evaluated.

5. CONCLUSION

Advocates for a new national testing policy abound. There are many true believers in the Victorian dogma, preached by proselytizers of the church of measurement-driven instruction, that salvation lies in the technology of a high-stakes system of exams. The lure of its promises are intoxicating. Precisely because of this atmosphere there

is a need to be reminded of the, "timeless precedence of 'thou shall *not*'over 'thou shalt' in ethics" (Jonas, 1984, p. 203). The technophiles speak for themselves all over the place but "a dissenting voice is sometimes needed to moderate the din made by the enthusiastic multitudes" (Postman, 1992, p. 5).

When policy makers consider mandating a national examination or system of examinations they need to be fully aware of the fact that a national test is a technology. The history of technology in general, and of testing in particular with its ironic turns, should be read as a cautionary tale. I believe that we do not yet know the answer to crucial technical and equity questions embedded in the creation of a high-stakes national examination system, nor have we estimated its potential impacts on individuals, schools, and the workplace. Given what I believe is at best an agnostic answer to important technical and equity questions, we should proceed with the utmost caution, examine claims, anticipate impacts, solve technical and equity problems, and put societal and institutional safeguards in place before mandating a high-stakes national examination system for all students.

References

Airasian, P. (1988). Measurement driven instruction: A closer look. Educational Measurement: Issues and Practice, 7(2), 6-10.

Airasian, P. (1993). Critical pedagogy and the realities of teaching. In H. J. Perkinson (Ed.), Teachers without goals students without purposes, (pp. 81-93). New York: McGraw-Hill, Inc.

Applebaum, H. (1992). The concept of work: Ancient, medieval, and modern. Albany, NY: State University of New York Press.

Baker, E. L., O'Neil, H. F., & Linn, R. L (1991). Policy and valididy prospects for performance-based assessment. Paper presented at the Annual Meeting of the American Psychological Association.

Basalla, G. (1988). The Evolution of technology. New York: Cambridge University Press.

Bell, R., & Grant, N. (1974). A Mythology of British Education. London: Panther.

Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A., & Tipton, S. M. (1985). Habits of the heart. Berkeley: University of California Press.

Bellah, R. N., Madsen, R., Sullivan, W. M., Swidler, A., & Tipton, S. M. (1991). The good society. New York: Alfred A. Knopf.

Berdyaev, N. (1972). Man and machine. In C. Mitcham & R. Mackey (Eds.), Philosophy and technology: Readings in the philophical problems of technology (pp. 203-213). London: The Free Press.

Bijker, W. E., & Law, J. (1992). General introduction. In W. E. Bijker & J. Law (Eds.), Shaping technology/Building society: Studies in sociotechnical change (pp. 1-14). Cambridge, MA: The MIT Press.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27(2), 165-174.

Boorstin, D. J. (1978). The Republic of technology. New York: Harper & Row, Publishers, Inc.

Borgmann, A. (1984). Technology and the character of contemporary life: A philosophical inquiry. Chicago: University of Chicago Press.

Cannell, J. J. (1987). Naturally normed elementary achievement testing in America's public schools: How all fifty states are above the national average. Daniels, WV: Friends for Education.

Cannell, J. J. (1989). The "Lake Wobegon" report: How public educators cheat on standardized achievement tests. Albuquerque, NM: Friends for Education.

Carroll, J. B. (1987). The National Assessment in reading: Are we misreading the findings? Phi Delta Kappan, 68(6), 424-430.

Carter, R. S. (1952). How valid are marks assigned by teachers? Journal of Educational Psychology, 43(218-228).

Commission on Chapter 1 (1992). Making schools work for children in poverty. Washington, DC: Education Commission of the States.

Darling-Hammond, L. (1991). The implications of testing policy for educational quality and equality. Phi Delta Kappan, 73(3), 220-225.

Deming, W. E. (1982). Out of the crisis. Cambridge, MA: The MIT Center for Advanced Engineering Study.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessment. Applied Measurement in Education, 4(4), 289-304.

Elam, S. M, & Gallup, A. M. (1989). 21st. Annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 71(1), 41-57.

Elam, S. M. (1990). 22nd. Annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 72(1), 41-53.

Ellul, J. (1964). The Technological Society (Wilkinson, John, Trans.). New York: Vintage Books.

Ellul, J. (1990). The technological bluff. Grand Rapids, MI: Williams B. Eerdmans Company.

Ellul, J. (1992). Technology and democracy. In L. Winner (Ed.), Democracy in a technological society (pp. 35-50). Boston: Kluwer Academic Publishers.

Elton, R. G. (1991). Return to essentials: Some reflections on the present state of historical study. Cambridge: Cambridge University Press.

Fielder, J. (1992). Autonomous technology, democracy, and the nimbys. In L. Winner (Ed.), Democracy in a technological society (pp. 105-121). Boston: Kluwer Academic Publishers.

Foucault, M. (1979). Discipline and punish: The birth of the prison. Harmondsworth, England: Penguin.

Frost, R. L. (1992). Mechanical dreams: Democracy and technological discourse in twentieth-century France. In L. Winner (Ed.), Democracy in a technological society (pp. 51-77). Boston: Kluwer Academic Publishers.

Gallup, A. (1986). 18th. Annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 68(1), 43-60.

Gallup, A. M, & Elam, S. M. (1988). 20th.Annual gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 70(1), 33-46.

Gallup, A. M. (1985). The Gallup poll of teachers' attitudes toward the public schools Part 2. Phi Delta Kappan, 66(5), 323-330.

Gallup, A. M., & Clark, D. L. (1987). The 19th annual Gallup poll of the public's attitudes towards the public schools. Phi Delta Kappan, 69(1), 17-30.

Gallup, G. (1978). 10th. Annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 60(1), 34-46.

Gallup, G. (1980). 12th. Annual Gallup poll of the public's attitudes toward the public schools. Phi Delta Kappan, 62, 33-46.

Gould, S. J. (1981). The mismeasure of man. New York: W. W. Norton & Company.

Gracian, B., S.J., (1653). The art of worldly wisdom (Jacobs, J, Trans.). (1936 ed.). London: Macmillan & Co.

Haertel, E. (1989). Student achievement tests as tools of educational policy: Practices and consequences. In B. Gifford (Ed.), Test Policy and Test Performance: Education, Language and Culture (pp. 25-50). Boston: Kluwer Academic Publishers.

Haladyna, T. M., Nolen, S. B., & Hass, N. S. (1989). Report to the Arizona legislature: Test score pollution. Pheonix, AZ: Arizona State University West Campus.

Hambleton, R., & Murphy, E. (1992). A psychometric perspective on authentic measurement. Applied Measurement in Education, 5(1), 1-16.

Haney, W., Madaus, G. F, & Lyons, R. (1993 forthcoming). The fractured market place for standardized testing. Boston: Kluwer Academic Publishers.

Holmes, E. G. A. (1911). What is and what might be: A study of education in general and elementary in particular. London: Constable.

Holt, M. (1993). The educational consequences of W. Edwards Deming. Phi Delta Kappan, 74(5), 382-388.

Hoskins, K. (1968). The examination, disciplinary power and rational schooling. History of Education, 8, 135-146.

Hottois, G. (1987). Technoscience: Nihilistic power versus a new ethical consciousness. In P. T. Durbin (Ed.), Technology and responsibility. (Official publication of the Society for Philosophy and Technology, 3 ed.) (pp. 69-84). Boston: D. Reidel Publishing Company.

Hughes, T. P. (1989). American genesis: A century of invention and technological enthusiasm. New York: Penguin Books.

Jaeger, R. M., (1991). Legislative perspectives on statewide testing: Goals, hopes, and desires. Phi Delta Kappan., 73(3), 239-242.

Jonas, H. (1984), The imperative of responsibility: In search of an ethics for the technological age. Chicago: University of Chicago Press.

Kermode, F. (1988, September 18). I am only equivocally Harold Brodkey. The New York Times Book Review, p. 3.

Koretz, D. M., Linn, R. L., Dunbar, S. B., & Shepard, L. A. (1991). The effects of high stakes testing on achievement: Preliminary findings about generalization across tests. Paper presented at the Annual meeting of the American Educational Research Association, Chicago, IL.

Koretz, D. M., Madaus, G. F., Haertel, E., & Beaton, A. (1992). Statement before the Subcommittee on Elementary, Secondary, and Vocational Education Committee on Education and Labor U.S. House of Representatives, February 19,1992

Kozol, J. (1991). Savage inequalities: Children in America's schools. New York: Crown Publishers, Inc.

Kreitzer, A. E., Haney, W., & Madaus, G. F. (1989). Competency testing and dropouts. In E. F. Lois Weis and Hugh G. Petrie (Eds.), Dropouts from school: Issues, dilemmas, and solutions, Part II (pp. 129-152). New York: State University of New York.

Kula, W. (1986). Measures and Men (Szreter, R. Trans.). Princenton, NJ: Princeton University Press.

LeMahieu, P. (1992). Personal communication. December 23,1992.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, perfromance-based assessment: Expectations and validation criteria. Evaluation Comment, (Winter, 1991-92), 2-9.

Linn, R. L., Graue, B., & Sanders, N. M. (1989). Comparing state and district test results to national norms: Interpretations of scoring "above the national average", Paper presented at the Annual meeting of the American Educational Research Association.

Lowrance, W. W. (1986). Modern science and human values. New York: Oxford University Press.

Madaus, G. F. (in press). A national testing system: Manna from above? A historical/technological perspective. Journal of Assessment in Education, 1(1).

Madaus, G. M. (1991). The effects of important tests on students: Implications for a national examination or system of examinations. Phi Delta Kappan, 73(3), 226-231.

Madaus, G. F, Haney, W., Newton, K. B, & Kreitzer, A. (1992). A proposal for a monitoring body for tests used in public policy. Paper presented at the Conference on The Evaluation of Test-based Educational Reforms. Executive Education, Babson College Wellesley, MA: Sponsored by Center for the Study of Testing Evaluation and Public Policy, Boston College and the RAND Institute on Education and Training.

Madaus, G., F., & Kellaghan, T. (1991a). Student examination systems in the European Community: Lessons for the United States Contractor Report submitted to the Office of Technology Assessment, United States Congress.

Madaus, G. F., & Kellaghan, T. (1991b). National testing: Lessons for America from Europe. Educational Leadership 49(3), 87-93.

Madaus, G. F., & Kellaghan, T. (1992). Curriculum evaluation and assessment. In P. W. Jackson (Ed.), Handbook of research on curriculum (pp. 119-154). New York: Macmillian.

Madaus, G., F., & Kellaghan, T. (1993). The British experience with 'Authentic' testing. Phi Delta Kappan, 74(6), 458-469.

Madaus, G. F., & Greanery, V. (1985). The Irish experience in competency testing: Implications for American education. American Journal of Education, 93(2), 268-294.

Madaus, G. F., & McDonagh, J. T. (1979). Minimum competency testing: Unexamined assumptions and unexplored negative outcomes. New Directions for Testing and Measurement, (3), 1-15.

Madaus, G. F., West, M. M., Harmon, M. C., Lomax, R. G., & Viator, K. (1992). The influence of testing on teaching math and science in grades 4-12 Report of a study funded by the National Science Foundation (No. SPA8954759). Boston: The Center for the Study of Testing, Evaluation and Educational Policy, Boston College.

McCandless, B. R., Roberts, A., & Starnes, T. (1972). Teacher marks, achievement test and scores, and aptitude relations with respect to social class, race and sex. Journal of Educational Psychology, 63, 153-159.

McLauglin, M. W., (1991). Test-based accountability as a reform strategy. Phi Delta Kappan, 73(3), 248-250.

McShane, D. (1989). Testimony presented at a hearing entitled The Effects of Testing on American Indians. Co-sponsored by the National Commission on Testing and Public Policy and the Native American Scholarship Fund, Inc., Albuquerque: University of New Mexico.

Merton, R. K. (1964). Foreword. In Ellul, J., The technological society (pp. v-viii). New York: Vintage Books.

Monroe, P. (1936). Conference on examinations. New York: Bureau of Publications, Teachers College, Columbia University.

National Commission on Testing and Public Policy (1990). From gatekeeper to gateway: transforming testing in America. Chestnut Hill, MA: National Commission on Testing and Public Policy, Boston College.

National Council on Education Standards and Testing (1992). Raising Standards for American Education: A Report to Congress, the Secretary of Education, the National Education Goals Panel, and the American People. Washington, D.C.: Author.

Nuttall, D. L. (1992). Performance assessment: The message from England. Educational Leadership, 49(8), 54-57.

Ong, W. J. SJ. (1971). Rhetoric, romance, and technology. Ithaca, NY: Cornell University Press.

Ong, W. J. SJ. (1977). Interfaces of the world: Studies in the evolution of consciousness and culture. Ithaca, NY: Cornell University Press.

Ong, W. J. (1982). Orality and literacy: The technologizing of the word. London: Routledge.

Pacey, A. (1989). The culture of technology. Cambridge, MA: MIT Press.

Popham, W. J. (1993). Circumventing the high costs of authentic assessment. Phi Delta Kappan, 74(6), 470-473.

Postman, N. (1992). Technoloply: the surrender of culture to technology. New York: Alfred A. Knopf.

Rapple, B. A. (1992). A Victorian experment in economic efficiency in education. Economic of Education Review 11(4), 301-316.

Reiser, S. J., (1978). Medicine and the reign of technology. Cambridge: Cambridge University Press.

Rotenstreich, N. (1972). Technology and politics. In C. Mitcham & R. Mackey (Eds.), Philosophy and technology: Readings in the philophical problems of technology (pp. 151-160). London: The Free Press.

Rothblatt, S. (1982). Failure in early nineteenth century Oxford and Cambridge. History of Education, 11(1), 1-21.

Sarason, S. (1990). The predictable failure of educational reform. San Francisco CA: Jossey-Bass.

Sclove, R. E. (1992). The nuts and bolts of democracy: Democratic theory and technological design. In L. Winner (Ed.), Democracy in a technological society (pp. 139-157). Boston: Kluwer Academic Publishers.

Sedwich, J. (1993). The complexity problem. Atlantic Monthly, 271(3), 96-104.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1991). Performance assessment in science. Applied Measurement in Education., 4(4), 347-362.

Shavelson, R. J., Baxter, G. P., & Pine, J. (1992). Performance assessments: Political Rhetoric and measurement reality. Educational Researcher, 21(4), 22-27.

Shepard, L. A. (1989a). Inflated test score gains: Is it old norms or teaching to the test. Paper presented at the Annual meeting of the American Educational Research Association.

Shepard, L. A. (1989b). Why we need better assessments. Educational Leadership, 46(7), 4-9.

Shepard, L. A. (1991a). Will national tests improve student learning? Phi Delta Kappan, 73(3), 232-238.

Shepard, L. A. (1991b). Negative policies for dealing with diversity: When does assessment and diagnosis turn into sorting and segregation? In E. Hiebert (Ed.), Literacy for a diverse society: Perspectives, practices, and policies. (pp. 331-352). New York: Teachers College Press.

Shrader-Frechette, K. (1987). Risk-cost- benefit analysis. In P. T. Durbin (Ed.), Technology and responsibility. (Official publication of the Society for Philosophy and Technology, 3 ed.) (pp. 343-357). Boston: D. Reidel Publishing Company.

Shrader-Frechette, K. (1992). Technology, bayesian policymaking, and democratic process. In L. Winner (Ed.), Democracy in a technological society (pp. 123-137). Boston: Kluwer Academic Publishers.

Simon, Y. R. (1983). Pursuit of happiness and lust for power in technological society. In C. Mitcham & R. Mackey (Eds.), Philosophy

and technology: Readings in the philosophical problems of technology (pp. 171-186). New York: Free Press.

Simpson, J. A., & Wiener, E. S. C. (Eds.). (1989). Oxford English Dictionary. Oxford: Oxford University Press.

Sizer, T. R. (1984). Horace's compromise: The dilemma of the American high school. Boston: Houghton Mifflin Company.

Stake, R. E. (1991). The teacher, standardized testing, and prospects of revolution. Phi Delta Kappan, 73(3), 243-247.

Stanley, M. (1978). The technological conscience: Survival and dignity in an age of expertise. Chicago: University of Chicago Press.

Staudenmaier, J. (1988). Technology and faith. Audio cassette. Kansas City: Credence Cassettes.

Staudenmaier, J. M. (1985). Technology's storytellers: Reweaving the human fabric. Cambridge MA: MIT Press.

Staudenmaier, J. M. (1989). U.S. technological style and the atrophe of civic commitment. In D. L. Gilpi (Ed.), Beyond individualism toward a retrieval of moral discourse in America (pp. 120-152). South Bend: Notre Dame Press.

Taylor, C. (1991). The ethics of authenticity. Cambridge, MA: Harvard University Press.

Wajcman, J. (1991). Feminism confronts technology. Cambridge, England: Polity Press.

Wallace, W. A., (Ed.). (in press). Ethics in modelling. London: Pergamon Press.

Wartofsky, M. W. (1992). Technology, power, and truth: political and epistemological reflections on the fourth revolution. In L. Winner (Ed.), Democracy in a technological society (pp. 15-34). Boston: Kluwer Academic Publishers.

Winner, L. (1977). Autonomous technology: Technic-out-of-control as a theme in political thought. Cambridge MA: MIT Press.

Winner, L. (1986). The Whale and the reactor: A search for limits in an age of high technology. Chicago: The University of Chicago Press.

Winner, L. (Ed.). (1992a). Democracy in a technological society (Official publication of the Society for Philosophy and Technology, 9th ed.). Boston: Kluwer Academic Publishers.

Winner, L. (1992b). Introduction. In L. Winner (Ed.), Democracy in a technological society (pp. 1-11). Boston: Kluwer Academic Publishers.

Wood, R. (1984). Observations on criterion referenced assessment. Paper prepared for a seminar held at the Senior Secondary Assessment Board of Australia.

End Notes

[1] I would like to thank Thomas Kellaghan of the Educational Research Centre St Patrick's College Dublin Ireland and Stacey Raczek and Ann Tan of Boston College for their generous comments and assistance.

[2] Current and proposed testing programs, the instruments themselves, and the arcane psychometric underpinning of scoring and reporting are all consistent with various ways of defining technology found in the literature. For example, technology is any complex of standardized means, special knowledge and activities, skills, methods, and procedures people use to attain a predetermined end in social, economic, administrative and educational institutions (Ellul, 1964, 1990; Lowrance, 1986; Winner, 1977). Testing fits this, and all other definitions of technology I have come across. For other definitions see (Basalla, 1988; Boorstin, 1978; Borgmann, 1984; Pacey, 1989; Staudenmaier, 1988, 1985; Staudenmaier, 1989; Winner, 1986)

[3] Ong (1977) offers one of the most compelling descriptions of the benefits of technology:

> Technology is important in the history of the word not merely exteriorly, as a kind of circulator of pre-existing materials, but interiorly, for it transforms what can be said and what is said. Since writing came into existence, the evolution of the word and the evolution of consciousness have been intimately tied in with technologies and technological developments. Indeed, all major advances in consciousness depend on technological transformations and implementations of the word. (Ong, 1977 p. 42)

[4] The famous historian of technology Thomas P. Hughes in a caution, apt to those who would put a new national high-stakes system of exams in place, reminds us that:

> ... "Americans need to fathom the depths of the technological society, to identify currents running more deeply than those conventionally associated with politics and economics. Americans [need to realize] that not only their remarkable achievements but many of their deep and persistent problems arise, in the name of order, system, and control, from the mechanization and systematization of life and from the sacrifice of the organic and spontaneous "(Hughes, 1989).

Another commentator describes this unawareness of technologies hidden effects this way, "It is perplexing as if for generations a family shared its home with a giant, orange, seizure-prone elephant, and yet never discussed-- somehow never noticed--the beast's presence and pervasive influence upon every facet of their lives" (Sclove, 1992, p. 139).

[5] For a description of how the current testing policies of the past decade were corrupted see for example (Cannell, 1987, 1989; Haertel, 1989; Haladyna, Nolen, & Hass, 1989; Kreitzer, Haney, & Madaus, 1989; Linn, Graue, & Sanders, 1989; Madaus & Kellaghan, 1992; Madaus & McDonagh, 1979; Madaus, 1991; Shepard, 1989a, 1989b,1991b; Madaus, West, Lomax, Harmon and Viator, 1992). For an example of how writing samples can be corrupted over time see (Madaus & Greanery, 1985)

End Notes Cont.

[6] I want to emphasis that I believe strongly that performance-based assessments in the hands of teachers, integrated in normal classroom routines, should be more useful to them for formative and diagnostic purposes than traditional standardized tests have ever been. But this is not my focus.

Writing in 1653 the Jesuit Balthsar Gracian observed that "Great judgment is needed to test that of another...Here one requires extraordinary care, deep observation, subtle discernment, and judicious decision (Gracian, 1653 p. 174-175) These virtues can only be fulfilled by classroom teachers; they are not characteristics of an impersonal bureaucratic testing program. Further, as I have stated elsewhere I believe that teachers, not assessments, must be the cornerstone of any systemic reform directed at improving our schools (Madaus, in press). Policy makers lose sight of the basic fact that, -- and I paraphrase slightly Parker Palmer here-- "the teacher is a mediator between the knower and the known, between the learner and the subject to be learned. A teacher, not some [test or performance assessment], is the living link in the epistemological chain" (Palmer, 1983, p 29-30). One of the reasons we mandated high-stakes testing systems is because we do not trust teachers. An excerpt from Sue Halpern's book *Migrations to Solitude* (Pantheon) recently published in the *Noted With Pleasure* section of the New York Times Book Review (3/15/92, p 31) captures much of what I think is wrong about the way we use tests in the policy sphere. She writes:

> Trust, which is a virtue, is also a habit, like prayer. It requires exercise. And just as no one can run five miles a day and cede the cardiovascular effects to someone else, no one can trust for us. Forgetting how to trust, we rely on random drug tests, honesty examinations, telephone monitoring and credit checks of prospective suitors.

In the Deming model of total quality management trust is a key element and our present forms of accountability only undercuts trust (Holt, 1993). While watching PBS recently, I was interested to hear W. Edwards Deming assert that whenever you manage by results, whether in industry or education, quality goes down.

It's a shame that our policy makers haven't put forward a national educational goal that states: "By the year 2000 America will have the best prepared, most highly respected teaching corps in the world."

[7] Winner tells us that:

> technological innovations are similar to legislative acts or political foundings that establish a framework for public order that will endure over many generations. For that reason the same careful attention one would give to the rules, roles, and relationships of politics must also be given to such things as the building of highways, the creation of television networks, and the tailoring of seemingly insignificant features on new machines (Winner, 1986 p. 26)

And I might add the same consequences and therefore the same considerations must be given to our public policy tests and testing programs.

[8] For some concrete examples see for example, (Basalla, 1988; Boorstin, 1978; Borgmann, 1984; Ellul, 1964; Pacey, 1989; Winner, 1977, 1986, 1992a;Berdyaev, 1972)

[9] See (Haney, Madaus, & Lyons, 1993 forthcoming) for a description of the many faceted testing industry.

[10] This shift in testing from a means to an end is an abiding problem in education and is wonderfully illustrated in the following observation from a nineteenth century British school inspector who observed first hand the negative effects of a high stakes

End Notes Cont.

testing program operating in England and Ireland which tied pupil examination results to teacher salaries:

> Whenever the outward standard of reality (examination results) has established itself at the expense of the inward, the ease with which worth (or what passes for such) can be measured is ever tending to become in itself the chief, if not sole, measure of worth. And in proportion as we tend to value the results of education for their measurableness, so we tend to undervalue and at last to ignore those results which are too intrinsically valuable to be measured (Holmes, 1911).

Bellah and associates (1991) point out that:

> Institutuions are forever subject to human corruption that comes from making means into ends.... Yet it is also true that ends ,and ends and means are subtly intertwined in human institutions and it is not always clear which is which. (p. 290)

[11] From a technological perspective equity is "based on the empirical reality of the confrontation of [people] and instruments, of the general call for these instruments, and the style brought about by their existence and use" (Rotenstreich, 1972 p. 154).

[12]For a detailed treatment of the issues surrounding the ethics of modelling see (Wallace, in press)

[13]For example, the use of the three parameter item response model by several commercial vendors in designing their elementary and secondary school achievement tests embodies value judgments and assumptions about the underlying ability level and guessing behavior of the test taker. However, most educators, students, and their parents have little or no knowledge of how these psychometric values and assumptions influence the derived scores of different students. In fact, most do not even know that the arcane three parameter model is even used at all.

[14] See for example Wajcman, (1991) for a discussion of how technology embodies male rather than female values and the need for a technology based on women's values.

[15] While I am focusing on how the technology of testing stacks the deck I am fully aware of the fact that the school system itself first unintentionally stacks the deck against certain populations which in turn exacerbates testing's influence. Sizer (1984) for example describes the different feel of schools enrolling poor children from the feel of schools enrolling more affluent students. The former are more hectic, pressured, less predictable, with more menace, and less arrogance and sense of expected entitlement. Also see Kozol (1991)

[16] See Postman (1992, p.125) for an excellent discussion of how the structure of a test question is not neutral; its form may ease an examinee's way or pose obstacles.

[17] Different cultural groups in American society may have different intellectual traditions which create different conceptions of reality than that tapped by our testing instruments. Consider the example of a young Ojibwa student tested by several professionals and classified as having certain learning and behavioral problems. In part, this classification was based upon his staring vacantly into space, completing tasks very slowly, and giving "non-reality based" responses to questions. As it turned out, the boy had a special relationship with his traditional Ojibwa grandfather, who encouraged his dreaming whether it occurred by day or by night and often discussed the nature of dreams with him. In Ojibwa thought and language, *ga-na-wa-bun-daw-ming*, which means seeing without feeling (objectivity), carries less value than *mu-zhi-tum-*

End Notes Cont.

ing, which means feeling what you do not see (subjectivity) .(McShane, 1989; National Commission on Testing and Public Policy, 1990) Our tests tend to promote, not just for Ojibwas but for all of us, the values of objectivity, the importance of factual knowledge and "right" answers, and rapid visible performance; in doing so, they devalue subjectivity, feelings, reflection, introspection, and discernment.

[18] Rapple (1992) has an excellent historical treatment of an earlier attempt at market driven education the Victorian payment by results scheme. Rapple ,while very sensitive to contextual, political, economic and cultural difference, compares present proposals for market driven education with the payment by results plan and concludes that similar negative outcomes associated with the 19th century effort will likely result. In regard to effects on teachers he concludes that school choice plans based primarily on exam results will further deprofessionalize teachers lessen their status, autonomy, effectiveness and true function as pedagogues.

[19]Bellah and his associates (1991) describe the place of the individual in this power mix through a quote from the anthropologist Mary Douglas "The most profound decisions about justice are not made by individuals as such, but by individuals thinking within and on behalf of institutions." (1991, p.13)

20 Thus, present rhetoric calls for a "voluntary," "national but not federal" system of examinations. These qualifiers are an implicit recognition of how testing is involved with larger structures of power in our society. Nonetheless, we already have a quasi federal, non voluntary, testing system operating in this country in Chapter 1. Further, a consideration of the history of NAEP shows that initial safeguards against federal intervention into state performance has become eroded with the trial state assessments.

[21] In testing this politicization takes one of its forms in mandating census testing of all students at a given grade level when sampling information because policy makers are acutely aware that all parents vote and want information on their child. Nonetheless, sampling would provide policy makers with accurate and cheaper accountability information

[22] On the latter point technocrats from both parties would use tests to centralize and control education, under the banner of competition and accountability while simultaneously proclaiming the need to get government out of our lives to de-centralize to de-regulate. In the economic sphere, Galbraith calls this dual mentality the approved contradiction (Stanley, 1978).

[23] Carroll (1987) in commenting on the NAEP reading results describes a problem with the measurement and the teaching of higher order thinking skills. He pointed out that many of the more difficult tasks used in the NAEP reading assessment -- those measuring so-called higher-order thinking -- resemble tasks found on tests of verbal and scholastic aptitude. Although he recognizes that it may be unpopular, he wonders if these exercises may be a measure of the verbal ability of the nation's youth, and if this is so, he points out that the research literature does not hold out much hope of improvement, given what we know about teaching a curriculum geared to higher-order skills. Sizer (1984) describes how these higher order skills might be taught and learned. However, he also notes that "some students are 'smarter' than others some of the time, but we don't know quite why, certainly not in a way that allows us easily to correct for it.' (p. 39). Airasian (1988, 1993) describes the difficulty of measurement driven instruction (MDI) aimed at higher level cognitive outcomes . He writes , "Given existing resources, teacher preparation, and school organization we are unable to teach

End Notes Cont.

processes such as application, critical thinking, analysis, and reasoning to most pupils" (1989, p. 10). Airasian (1993) expands on the difficulties of critical pedagogy as schools are presently staffed and constituted. All of this points to the critical necessary of better pre- and in-service teacher preparation before embarking on a mandated program of high-stakes assessments supposedly geared to higher-order thinking skills.

[24] Technical experts are not exclusively testing experts in this latest round of test driven reform. In fact, present test driven reform is being pushed by cognitive psychologists, educational policy analysts and curriculum experts who often cavalierly dismiss psychometric concerns as a minor problem, distraction, or nuisance.

[25] Hottois claims that "Much of what comes together under the standard of 'participation' (or similarly, of 'technological culture') is no more thanmystification, sand in the eyes, an ideological trick intended to allow technological society to preserve a deomocratic [sic] mask" (p.79). Fielder in his treatment of the NIMBY phenomenon points out that while it is true that citizens are not well-informed about technologically sophisticated issues being debated in legislatures, they have repeatedly shown that they can become knowledgeable about specific issues that deeply concern them. (Fielder, 1992)

[26] The annual Gallup Poll is often cited as evidence that citizens favor the establishment of a national exam system. *The Annual Gallup Poll of the Public's Attitude Toward The Public Schools* offers a gauge on how attitudes toward a national test and a national curriculum have evolved over the last two decades. The data from the various Gallup polls would indicate support for a national test -- even one for graduation decisions -- and a national curriculum of some kind (Elam & Gallup, 1989; Elam, 1990; Gallup, 1986; Gallup & Elam, 1988; Gallup, 1985; Gallup & Clark, 1987; Gallup, 1978; Gallup, 1980; Madaus & Kellaghan, 1991). (See Madaus & Kellaghan, (1991a) for a full treatment of the Gallup data). Madaus and Kellaghan recommend, that before concluding that any national test or curriculum at any grade, for any purpose would meet with public favor, questions would need to be more sharply drawn. For one thing it would be wise to tease out the tension between local control on the one hand and a national test and curriculum on the other. Respondents were never faced with a direct confrontation between the two concepts and most likely never considered the effects of a national test or curriculum on the cherished concept of local control. It would be wise to also ask:

- what subjects, content and levels would be included in a national curriculum.
- what the adjective "national" means, i.e., commercial, private, quasi-governmental, governmental;
- what such tests would be used for, i.e., promotion (so-called gates tests), graduation, certification of the successful completion of a given level of schooling or a given type of schooling (New York State Regents), selection (CEEB), school accountability(NAEP) etc.;
- what subjects/content/skills/abilities the tests would and would not measure and at what grade(s) the tests would be administered.

Further, in light of the strong sentiments in favor of local control, and that professional educators -- not state or federal government officials -- should set standards, questions about the sponsorship, control, operation and funding of a national test and national curriculum need to be asked. Finally, it would be wise to investigate public opinion concerning the variety and variations of national exams in European countries.

End Notes Cont.

[27] Forst (1992) points out that choices within technical systems at one time tend to prestructure later options, and how the social dynamics associated with specific systems bequeath legacies.

[28] Ong (1982) reminds us that written examination questions came into general use in the west thousands of years after the invention of writing. Classical Latin has no word for an 'examination' such as we take today and try to pass in schools. "Until the past few generations in the west, and still in perhaps most of the world today, academic practice has demand that students in class 'recite', that is feed back orally to the teacher statements (formulas--the oral heritage) that they had memorized from classroom instruction or from textbooks" (p. 56)

[29] When the data were broken out by number of Children in Chapter 1 the same results emerged.

[30] I would like to acknowledge the work of Stacey Raczek of Boston College in this review.

[31] We do know that males perform better than females on multiple choice tests compared to their performance on free-response examination questions (Bolger & Kellaghan, 1990). There is also evidence that teachers assign higher grades to girls than to boys see for example (Carter, 1952; McCandless, Roberts, & Starnes, 1972)

[32] Commenting on the introduction of new testing procedures in South Australia Wood offers this sound advice:

> Often when a call is made to replace X with Y it is conveniently forgotten that X has usually been the subject of much discussion and reflection over the years, whatever its shortcomings now appear to be, and that Y is untried and unexamined and, quite conceivably, is floated on a tide of euphoric expectation which is quite unrealistic. Sometimes it is better to patch up X and spend more time developing Y (Wood, 1984 p2.)

[33] Jonas (1984) reminds us once again of the autonomous nature of technology. He writes, "developments set in motion by technological acts with short-term aims tend to make themselves independent, that is, to gather their own compulsive dynamics, an automotive momentum, by which they become not only ... irreversible but also forward-pushing and thus overtake the wishes and plans of the initiators." (p 32)

[34] The following graphically illustrates how policy makers viewed the power of an important external examination to bring about changes, without directly dealing with either curriculum or the instructional-delivery system. The Prime Minister of Ireland, Eamon DeValera, arguing in Parliament for a system of certification examinations at the end of primary school, said in 1941:

> But if we want to see that a certain standard is reached and we are paying the money, we have the right to see that something is secured for that money. The ordinary way to test it and try to help the whole educational system is by arranging our tests in such a way that they will work in a direction we want (Dáil Éireann, 1941, col 119)

This example is not peculiar to its time or place. The sentiments which they reflect can also be found in contemporary America. The recent wave of reform reports contain proposals dealing with testing that embrace a strong bureaucratic belief in the power of external tests to direct the efforts of those in education.

End Notes Cont.

[35] This 1936 quote from an international conference on public examinations captures this paradox: "Gentlemen, the experience in our country in attempting to solve our difficulties has been that as soon as we think we have solved one problem we almost invariably find that by the solution we have created another" (Monroe, 1936 p. 115).

[36]For a discusssion of the benefits and downsides associated with the rise of individualism see (Bellah, Madsen, Sullivan, Swidler, & Tipton, 1985; Bellah, Madsen, Sullivan, Swidler, & Tipton, 1991; Taylor, 1991)

[37] This move of course predates by many centuries the Enlightenment and originated in China with the institution of its civil service exams. In the Western world civil service exams appeared in the 19th century.

[38] Another important aspect of technology's power in the political realm is the fact that it is one of the processes that increases and intensifies the bureaucratic character of modern states. According to Ellul technique is the paradigm of action furnishing organizations and process both its ontology and its underlying logic (Ellul, 1992). For a further discusssion of this fact see for example, (Rotenstreich, 1972) It could be argued that high-stakes testing programs are an example of this phenomenon.

[39] In the 16th century the foundation of the factory model of schooling was laid by Peter Ramus (1515-1572) and his followers through the introduction of the Ramus method of education which had as its controlling concept a method that Ong describes as a kind of intellectual commercialism. Ong goes on to point out that the Ramist approach to rhetoric made "it possible to think of knowledge itself in terms of "intake" and "output" and "consumption" -- terms which were not familiar to the commercial world in Ramus' day..but which [did] refer to realities present within that world" (Ong, 1971 p. 174). This commercial tendency in instruction, and in examining was aided in the late 18th century when William Farish invented the quantitative mark to individual questions on newly introduced written exams in the British universities. The combination of students responding in writing to a set of standardized questions, together with the ability to rank performance based on quantitative marks, led to a common set of questions defining a common curriculum. As Hoskins observed, the quantitative symbol of the mark had assumed supreme significance, and a generation later the questions posed assumed a factual "right" answer rather than being judge qualitatively. "Narrow specialization, and examination based on the principle of *testable knowledge* became the new parameters of undergraduate education...and with it a new intellectualist ideal, what we now call proficiency or the acquisition of skills" (Hoskins, 1968 p. 145, emphasis added) Factual knowledge and technical competence in specific subject areas rather than rhetorical style across the entire curriculum, assumed prominence; this in turn drastically modified what and how things were and learned. The combination of writing and quantitative scoring had tremendous bureaucratic potential. The quantitative score greatly increased the power of those using results to objectify examinees, document their attainment, compare and contrast them, and combine individual data to describe groups. Further, the introduction of a quantitative score might well be considered the first step in the development of the field of psychometrics, which is so dependent on the total score on a test.

[40] One historian reports that when the examination in England became a device to induce students to work hard, that is when students saw them as disciplinary tools, a world of student slang based on negative examination experiences sprang up. Many words for failing, cheating, and for survival strategies came into the vernacular (Rothblatt, 1982).

End Notes Cont.

[41] Madaus and Kellaghan have noted elsewhere that "Dedication to this credo, became for many an unswerving secular faith rivaling any religion. The symbolism of "scientific" or "objective" accuracy ascribed to testing has always been extremely important when tests are used for bureaucratic or administrative purposes. The potent symbolism of science and objectivity associated with testing shrouds testing in a pseudo-scientific myth, and gives test scores almost magical qualities particularly in the minds of lay persons (Bell & Grant, 1974) Unfortunately this symbolism and myth shifts the focus of much of the discussion about what is wrong with schools, or why certain groups perform poorly on tests, away from the critiques of the curriculum and the delivery of instruction, to acrimonious arguments for or against testing."

[42] A new modality of power was born in which each individual -- who Foucault describes as calculable man -- "receives as his status his own individuality, and in which he is linked by his status to the features, the measurements, the gaps, the 'marks' that characterize him and make him a case" (Foucault, 1979 p. 192).

[43] Harold Brodkey provides an example of the seductive, defining power of the IQ testing technology:

> I did well in school and seemed to be peculiarly able to learn what the teacher said -- I never mastered a subject, though -- and there was the idiotic testimony of those peculiar witnesses, IQ tests: those scores invented me. Those scores were a decisive piece of destiny in that they affected the way people treated you and regarded you; they determined your authority; and if you spoke oddly, they argued in favor of your sanity. But it was as easy to say and there was much evidence that I was stupid, in every way or, as my mother said in exasperation, 'in the ways that count.' (Kermode, 1988 p.3)

The experience of minority children offers a powerful example of the other side of the IQ coin from that provided by the Brodkey passage. When these children score low on an Anglo-oriented English language IQ test, they are labeled mentally deficient despite their different cultural and linguistic backgrounds.

[44] See Hottois (1987) for a discussion of the problem of the ideological component ofcontemporary science and technology.

[45] See Larry P. v Riles, 343 F.Supp. 1306 (n.d. Cal. 1972) Preliminary injunction granted, aff'd 502 F.2d 963 (9th Cir. 1974); 495 F.Supp. 926 (n.d. Cal. 1979), aff'd in part and rev'd in part, 793 F.2d 969 (9th Cir. 1984), 1986-87 EHLR Dec. 558:141 (N.D. Cal. 1986) (remedial decree modified)

[46] For a treatment of the effects on past testing policy on Chapter 1 see Commission on Chapter 1, 1992 . This report also proposed a radical modification of testing policies. I have presented my dissent to aspects of this proposed policy in an Appendix. Basically, I feel we do not yet know enough about proposed alternatives to make them the backbone of our largest federal program in education.

[47]The Oxford English Dictionary gives as one meaning of fair: Affording an equal chance of success; not unduly favourable or adverase to either side. Phase, *a fair field (and no favour)* P. 671 emphasis in original (Simpson & Wiener, 1989).

[48] Part of the problem in designing a test is the question of who is the user. User centered design makes a lot of sense for things that are used by different users in much

End Notes Cont.

the same way (Sedwich, 1993). However, this is not the case in testing where different users use the same test for quite different purposes.

[49] Cf footnote 15

[50] The Center for the Study of Testing, Evaluation and Educational Policy at Boston College under a grant from the John D. and Catherine C. Mac Arthur Foundation, The Pew Charitable Trust and the Ford Foundation is researching ways to use sampling techniques and new assessments to get rich information on school level achievement The project is called the Urban District Assessment Consortium (UDAC) and involves 11 urban school districts.

[51] Edward Haertel presented an excellent taxonomy of technical issues associated with alternative assessment used for individual decisions at a Conference sponsored by the Pew Charitable Trusts held in Phoenix AZ in July 1992. .

[52] Interestingly, it is the multiple choice item that is most widely employed in the external testing programs of our arch economic rival, Japan, to which we are often compared unfavorably by some of the same advocates of alternative assessment.

[53] Not that there aren't problems associated with these options.

[54] Most European countries had until after World War II what were called Primary Certificate exams given at the end of primary school to certify the successful completion of that level of education. For a discussion of how they operated and why they were eliminated see (Madaus & Greanery, 1985) The only European country testing primary children as part of national policy is Great Britain, hardly the economic competitor we wish to emulate. For problems associated with that system see (Madaus & Kellaghan, 1993; Nuttall, 1992)

[55] At least one proposal, that of the National Council on Education Standards and Testing (NCEST) would create a new entity, the National Education Standards and Assessment Council (NESAC), to provide "quality assurance" (The National Council on Education Standards and Testing, 1992). Commenting on this proposal, Koretz, Madaus, Haetrel and Beaton (1992) argued:

> If NESAC is established as proposed by the NCEST, however, its certification of assessments we feel would be a sham. Discharging its responsibilities would require substantial substantive and technical expertise, but the recommendations do not call for the appointment of even a single individual with expertise in measurement or evaluation. NESAC would also lack needed independence; its members would be appointed by the new National Education Goals Panel, and certifications would be made jointly by both organizations Equally important, the NCEST recommendations show a thorough misunderstanding of what is needed to validate tests and monitor their effects. A test cannot be validated by asking a group of individuals to examine its content, as the NCEST report implies. Moreover, validation is an ongoing process, not a one-time effort. To validate a test requires substantial empirical research, and the NESAC model does not make provisions for commissioning, funding, or using the needed investigations (Koretz, Madaus, Haertel, & Beaton, 1992).

[56] The current lack of public scrutiny of high-stakes tests is well illustrated by the experience of John Anderson, an English teacher at Cascade High School in Bell